Characterizing young children's everyday activities using video question-answering models

Tarun Sepuri*[1], Khai Loong Aw*[2], Alvin W.M. Tan*[2],
Robert Z. Sparks[2], Virginia A. Marchman[2], Michael C. Frank[2], Bria Long[1]
[1]University of California San Diego, San Diego, CA, 92093
[2]Stanford University, Stanford, CA, 94305
tsepuri@ucsd.edu
{khaiaw, tanawm, bsparks, marchman, mcfrank}@stanford.edu
brlong@ucsd.edu

Abstract

Children are remarkably efficient learners compared to our most advanced computational models of learning. One key difference is that children seem to leverage regularities in the activities (e.g., *eating*) in which they participate to learn about words or objects (e.g., "pomegranate"), even under skewed, long-tailed distributions. While everyday activities have long been theorized to be important as supports for children's learning, our understanding of the types, frequencies, and rhythms of these activities has been out of reach due to both a lack of naturalistic video datasets and the necessity for manual annotations. Here, we use the recent release of a large, egocentric dataset of children's everyday experience (BabyView) (N=31 children, N=868 hours) and capitalize on innovations in video question-answering (VideoQA) models to quantify the *what* and *where* of children's everyday experiences. Using these models, we classify both the activities (e.g., *eating, dancing, exploring*) and physical locations (e.g., *living room, garage*) in the infant view and generate natural-language descriptions for contiguous 10-second videos across the entire dataset. Notably, we find that (a) some activities and locations occur much more frequently than others, yet (b) there is wide variation across children. Moreover, (c) activities and locations exhibit structured transition probabilities (e.g., *cooking* often precedes *eating*), and (d) may decompose into distinct sub-clusters (e.g., different subtypes of *reading*). Compared with prior work analyzing static image content, our work highlights the advances possible by using VideoQA models to analyze the dynamic nature of children's experiences. Our results provide a better understanding of children's learning input in everyday contexts, informing developmentally-inspired models of early learning and cognitive development.

1 Introduction

Humans develop sophisticated visual-cognitive [1]–[3], linguistic [4], and motor capacities [5] early in childhood. Substantial research has sought to uncover the learning algorithms that give rise to these capacities by building computational models of human learning in these domains [6]–[14]. While leading models excel when trained on large, diverse datasets [15]–[17], they degrade markedly when trained on the same input that children receive [18]–[24]. Children are remarkably efficient learners relative to our best computational models of learning. What makes learning from children's everyday experiences challenging for models, but easy for children? One idea is that children may efficiently learn about words, actions, or objects by leveraging regularities in the activities in which they often

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Data on the Brain & Mind.

participate (e.g., *mealtime*) [25], [26]. For example, if the same sets of objects and labels appear frequently in a given context and location, then a new infrequent object – and its corresponding word – may be particularly salient and easier to learn given this stable context [27]. Children use known, highly frequent words to bootstrap learning of unknown, less frequent words: in fact, children and adults learn *more* efficiently from skewed, long-tailed distributions of both visual and linguistic inputs than from uniform distributions [28], [29]. Naturalistic data analyses also highlight the importance of *activity contexts*: in dense home video recordings from one child, words that were heard in more spatially-distinct contexts tended to be learned earlier in development [30] (see also [31]).

Activities are an important source of structure for children's everyday experiences – they differ along many social, linguistic, and pragmatic dimensions and in the affordances that they offer and therefore, in the learning opportunities that they provide [30], [32]–[36]. For example, book sharing typically consists of more lexically-diverse caregiver speech and more referential language than other types of activities (e.g., play) [33]–[36] and more structured activities tend to elicit more talk and more rare words from adult caregivers, e.g., mealtimes [33], [37]. Moreover, different kinds of activities vary in how often they occur and how long they last [38]. Finally, when participating in an activity, children are embedded in an environment that contains activity-specific speech [30], [35] and activity-specific visual information [26].

Thus, an understanding of the types, frequencies, and rhythms of everyday activities – how they relate to the objects and words that children experience – is a necessary step towards building ecologically-grounded, child-like models of early learning. At present, with a few exceptions [30], [39], [40], our understanding of these regularities has been limited to relatively small datasets coupled with manual activity coding and/or transcription. Further, existing activity and location annotation schemas are usually restricted to a small number of activities, usually owing to limited dataset size and the need for extensive manual annotations [35].

Here, we develop new methods for assessing the *what* and *where* of infants everyday experiences by leveraging recent innovations in video question-answering (VideoQA) models. We introduce a VideoQA model pipeline using `VideoLLaMA 3` [41] for the extraction of activities and locations present in a large dataset of children's naturalistic egocentric experiences [23] (N=868 hours of videos). We use a data-driven approach in order to quantify the activities and locations present in these videos. Specifically, we create a pipeline that automates the data-driven detection of activities and locations in contiguous 10-second chunks over the entire dataset and that produces semantically-coherent descriptions of these video segments. We then integrate these annotations with data on the estimated head-motion of each child from the accelerometer/gyroscope in the camera. Together, this information allows us to analyze the frequency and consistency of children's everyday activities in a large, longitudinal video dataset.

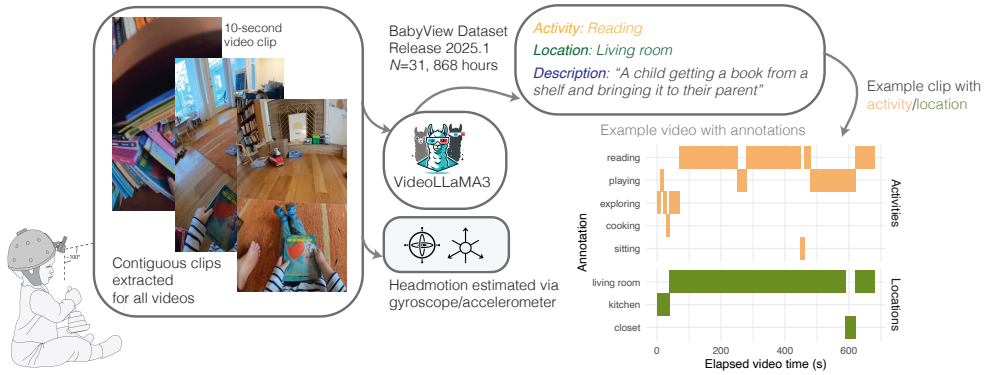

Figure 1: **Processing pipeline overview.** An automated pipeline for detecting the activities and locations in infant view using a video question answering model [41].

Figure 2: **Example frames from detected activities.** Example frames highlight that some activities in the infant view were child-centered (e.g., *playing*, *reading*) while others were detections of adult-centered activities that children mostly observed but were not engaged in directly (e.g., *cooking*).

2 Materials and Methods

2.1 Dataset

We analyze video recordings from a high-resolution head-mounted camera adapted for infants and children; the camera is a GoPro Hero Bones camera attached to a child-safety helmet, see [23]. In the 2025.1 release of the dataset, N=31 families recorded longitudinal data for a total of 868 hours across all children. These videos are accompanied by accelerometer/gyroscope data used to estimate children's own head-motion [42]–[44]. We extracted continuous, 10-second clips from all videos, for a total of N=312,485 short video clips.

2.2 Model annotation procedure

Annotation schema. With the intent of creating an annotation schema, we first prompted a VideoQA model, `VideoLLaMA 3` [41], with 10-second clips from the dataset. We acknowledge that 10 seconds is an arbitrary choice – and certain activities may be indeed longer or shorter than 10 seconds – but we anticipated that 10 seconds would be long enough to capture both transient and more sustained activities (e.g., *walking*). We first prompted the model to identify what activities and locations were present within each 10-second clip using a third of the dataset (N=293.3 hours), without constraining the model to a particular set of activities or locations. This provided a data-driven approach to identify what activities and locations may be present. Specifically, our prompt was: "This is a video from the point-of-view of a camera mounted on a child's head. Respond strictly only in this format with both keys and values: Location: … || Activity: …" Next, we then used these responses, as well as activity classifications from prior work [40], [45], to generate an annotation schema that included candidate activities and locations that would adequately capture those in the dataset.

Model annotation procedure. The 10-second clips were then presented sequentially to the VideoQA model, along with the revised annotation schema, with the addition of a catch-all location/activity category (*other*). The new prompt took the form "This video is recorded from the point-of-view of a child, with a camera mounted on the child's head. Respond strictly only in this format with both keys and values: Location: <balcony/bathroom/...> || Activity: <being held/cleaning/...> || Video description: ... ||" (exact prompt wording in Appendix A.1). The model gave three outputs for each video: an activity, a location, and a long-form (M=42.79 words) description of the video. Response generation was constrained to prefer tokens that fit our annotation schema. If the model generated a long-form description that was too short (less than 10 characters) or too long (greater than 550 characters), or an activity or location was not in our annotation schema, we re-ran each prompt a maximum of five times. In total, the VideoQA model could not generate a response for 1236 clips, representing 0.4% of generations. A manual examination of these clips suggested that they primarily consisted of dark rooms, camera pans to the wall, or zoomed-in play sessions where the model could not easily discern a location or activity. We ran the model on 8 NVIDIA A40 GPUs,

in parallel, in 122 hours. A schematic of our overall processing pipeline is shown in Figure 1. All of our code and processed data can be found at https://osf.io/ch56j/.

2.2.1 Annotation validation

Three trained coders each annotated the same 100 video clips. We asked coders to rate the quality of the linguistic descriptions of each video clip on the scale of 1–5. Annotators agreed to a moderate degree (Krippendorff's alpha (α)=0.52). In general, the manual coding indicated that the automated descriptions were reasonably correct, on average, across all three coders (M=3.27, SD=1.37). We also asked coders to list the activity and location in view using our schema. Specifically, we asked that coders annotate the activity that the child was seeing (if present) and not doing, to recover ground-truth annotations that were aligned with the VideoQA model. Overall, we found that labeling a single activity for a ten-second video clip was a relatively challenging task, with some disagreement between trained coders. There was moderate inter-rater agreement for the coders' location annotations (α=0.57, mean agreement=63.33%), while agreement between activity annotations was poorer (α=0.44, mean agreement=45.43%). Comparisons with the model annotations recapitulated this pattern (location precision=0.61, activity precision=0.41; precision values averaged across coders).

A manual inspection of a subset of the clips by authors suggested that a key reason why the activity detection task is non-trivial for both human raters and the VideoQA model is that multiple activities may co-occur even within a single clip (e.g., *reading* and *sitting*). To test this hypothesis, we asked annotators to additionally list all the activities that were ongoing in each clip, and not just the most salient one (activity count M=1.65, SD=0.68). Upon including these activities as alternative ground truths, the average activity precision increased to 0.51, and increased further to 0.67 when collapsing across annotators, providing credence to the hypothesis that some of the task ambiguity stems from the complex ways in which activities co-occur (see Appendix A.2 for more information).

2.3 Head-motion data: Accelerometer and gyroscope

The BabyView device includes two motion sensors: an *accelerometer* and a *gyroscope*, which capture two complementary kinds of movement. The accelerometer responds to translation (change in position); as the device also responds to the acceleration due to gravity, we subtracted the device's gravity estimate from the raw accelerometer signal, leaving linear acceleration due to actual movement. The gyroscope responds to rotation (change in orientation); we computed overall rotational activity by taking the Euclidean (ℓ_2) norm of angular velocity across the three axes of rotation. See Appendix A.3 for further details. To compare movement and head posture across annotated activities and locations, we processed the raw sensor streams into simple, orientation-robust features and summarized how these features varied by category.

3 Results

Prevalence and co-occurrence of activities and locations. We first examined the prevalence of the activities detected by our pipeline in the dataset, shown in Figure 3A for the eight most frequent non-postural activities (see Appendix A.4 for more information). These data were recorded in families' homes when it was convenient for families to record [23] and so, as expected, we found that the dominant activity identified was playing, and that most of these recordings took place in the living room (see Figure 8 in Appendix A.4). More broadly, we found a skewed distribution of the activities and locations in the infant view: some activities (e.g., *eating, playing*) and locations (e.g., *living room, kitchen*) were much more prevalent than others (e.g., *gardening* and *storage room*). When we examined the conjunction of activities within different locations, shown in Figure 4, overall, we found evidence for the face validity of our annotation schema: gardening tended to occur in the garden and cooking tended to occur in the kitchen. However, other activities (e.g., *playing*, *exploring*) were likely to occur in several different locations. Notably, although activity and location classifications were generated using the same prompt, they were not explicitly dependent on one another.

Individual variability across activities. We next examined how the proportion of identified activities differed across children, as shown in Figure 3A. We found variability in the distribution of time that children spent in different activities. Some of these differences may be attributable to sampling variation, with some caregivers using the head-mounted camera during more mealtime

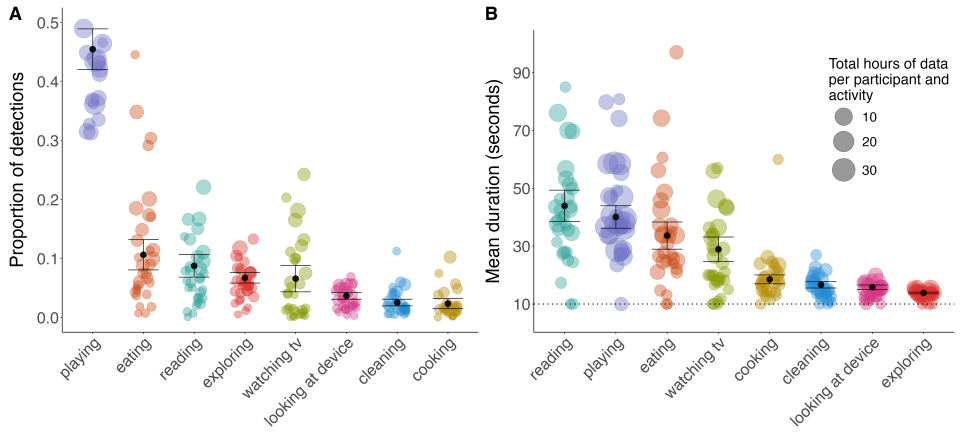

Figure 3: **Proportion and duration of activities.** (A) Average proportion of activities by participant. (B) Average durations of activities by participant. When a participant did not engage in a given activity, values were treated as missing (NA). The 8 most frequent non-postural activities are plotted (see Appendix A.4). Averages are weighted by the total number of hours a participant contributed to the dataset. Error bars are 95% CIs.

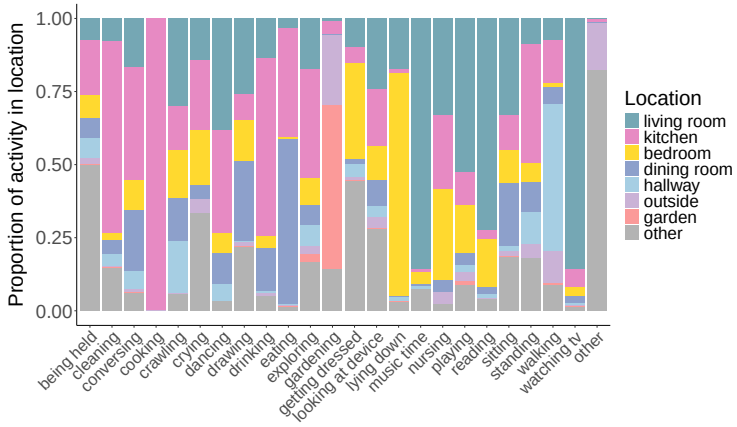

Figure 4: **Proportion of locations within activities.** Distribution of locations where each activity occurred, for the 6 most frequent locations and *garden*. Locations included *garden* to highlight the face validity of our annotation schema: gardening tended to occur in the garden. All other locations are collapsed into an eighth category *other*.

episodes. However, other differences appear to be grounded in variations in routines, given the broad variability in child-centered activities like *reading* and *watching TV*, for example. Some participants did not watch any television at all and others watched substantial amounts. In contrast, nearly all children engaged in substantial amounts of *playing*. While the present analyses do not yet systematically quantify age-related differences, we find qualitative evidence for developmental trends: for example, younger participants (5-11 month olds) had more *crawling* detections, while some older participants were more likely to be *watching TV*.

Temporal variability across activities. Next, we examined the mean duration of individual activities. 10-second activity detections are embedded within longer recordings in the dataset, with each recording lasting around 500 seconds on average (M=510.77, SD=383.10). For these analyses, we calculated the length of contiguous streams of detections (e.g., 12 clips of contiguous *playing* detections would be 120 seconds of playing) for each activity in each participant. These values are plotted in Figure 3B. We found that *reading* was the activity with the longest mean duration, followed by *playing*, *eating*, and *watching TV*. However, activity durations were only computed when the activity was detected in contiguous clips and therefore, are likely underestimating the duration

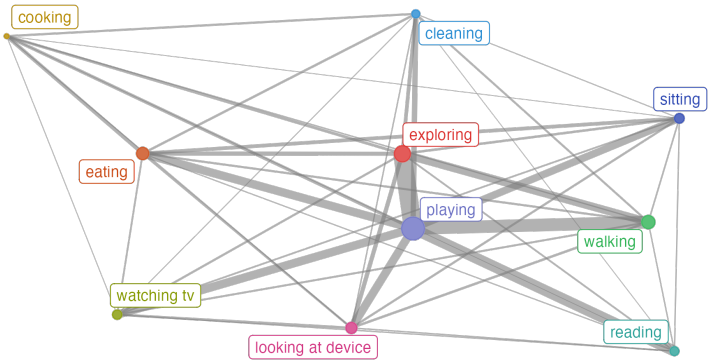

Figure 5: **Transition network over activities.** The 10 most frequent activities are shown, represented by nodes, sized by frequency. Edges represent bidirectional transitions between two activities, weighted by transition frequency. The network layout was generated using the Fruchterman–Reingold algorithm.

of certain activities (e.g., mealtime) when they are composed of micro-level activities (e.g., *eating, drinking, conversing*) that are separated into distinct detections.

Temporal sequences across activities. Temporal sequences across activities can also help to scaffold children's experiences – for example, cooking often precedes eating, and such transition regularities may allow children to reason about causality and predict subsequent experiences. To understand the broader activity macrostructure, we investigated the transition frequencies between activities. We first smoothed over activity annotations to reduce spurious transitions due to misclassified clips, such that each clip was annotated with the modal activity in a $\pm$10-second window around the clip (i.e., the most common activity across the n-1, n, and n+1 clips). We then tabulated transitions between different activities, regardless of the direction of transitions. The resultant network graph is shown in Figure 5. The transition frequencies among the activities reveal that activities are not randomly distributed, but follow transitional regularities; for example, *walking* is often temporally contiguous with *exploring* and *playing* but not *eating* or *cleaning*, and *cooking* often co-occurs with *eating* but not *watching TV*. Children's experiences are characterized not only by within-activity regularities, but between-activity regularities as well.

Content variability across activities. We next examined variation in the content of the detected activities by analyzing their descriptions. We embedded the descriptions using the Sentence Transformers model `all-MiniLM-L6-v2` into 384-dimensional embeddings, and visualized the embeddings using 2-dimensional t-SNE, as shown in Figure 6. The t-SNE plot recovered activity-level structure, with clusters emerging for different activity contexts; these clusters were also corroborated by random visual inspection of the corresponding clips, as shown in Figure 2. In addition, this visualization demonstrates the diversity of the video clips – for example, the clips of *playing* (the most frequent activity) span a wide region, reflecting the large variation in visual settings that can constitute play. Of course, *playing* likely consists of various subtypes of play – such as play with or without toys and with or without social partners, all of which present different affordances and learning opportunities. Figure 6 highlights descriptions from three different example *play* clips; quantifying these subtypes of play remains an avenue for future work.

However, *reading* was the activity that had the clearest subtypes: clips related to reading formed approximately three large clusters (see Figure 9 in Appendix A.5). The left cluster included clips in which the child was sitting on a caregiver's lap during shared book reading or by themselves with a book, with the book close to the child, near the bottom of their visual field. In contrast, the bottom cluster instead had clips in which the caregiver was located some distance from the child, holding up the book and reading from it. The center cluster consisted of more ambiguous clips in which books or paper were present but the observed activity was less clearly identifiable as an instance of shared book reading. Other activities also showed some substructure (e.g., *cleaning, watching TV*); investigating the organizational principles underlying these structures may be an interesting line of future research, the results of which will help to further refine the annotation schema.

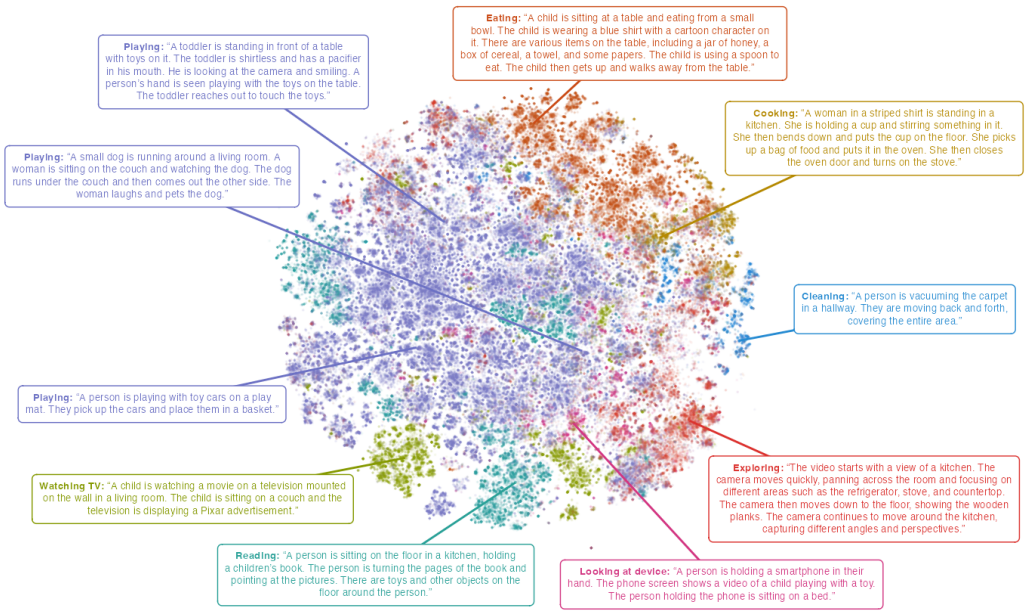

Figure 6: **t-SNE visualization of video description embeddings.** The 8 most frequent non-postural activities (see Appendix A.4), colored by activity. Text labels provide example model-generated descriptions and are selected from the same 10-second videos selected for the example frames shown in Figure 2.

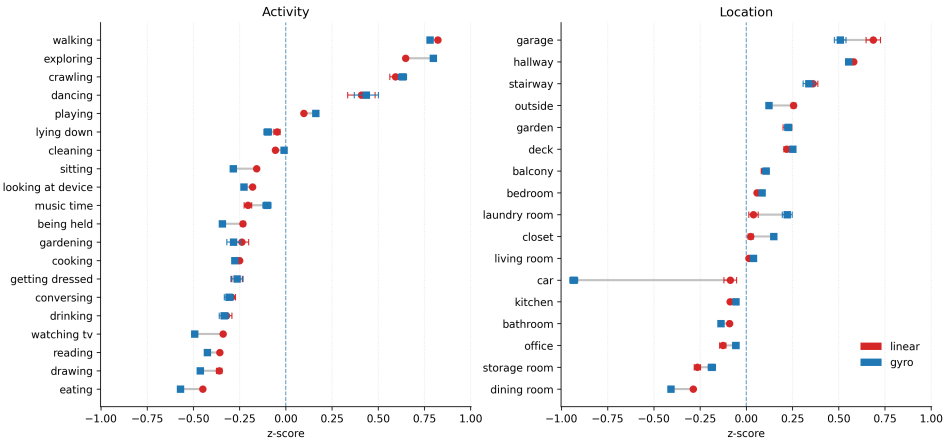

Figure 7: **Variability in head motion data by activity and location.** Magnitudes (z-scored) of total linear acceleration (red) and gyroscope rotation speed (blue), by activity (left) and location (right).

Head-motion variability across activities. Finally, we examined how children's motor experience varied across these activities. We expected that some activities (e.g., *walking*, *crawling*, *exploring*, *dancing*) would be more likely to involve children moving around on their own, while other activities would be more likely to involve children observing the world or focusing on specific actions with their hands. We found that some activities and locations tended to have more or less experienced head motion: for example, *crawling* and *walking* in the stairway, were characterized by higher than average head motion, and *eating* in the dining room, was characterized by lower than average head motion. Thus, these results offer both convergent validity for our main annotation schema and a new analytic lens for examining how children's own locomotive abilities intersect with their everyday experiences.

4 General Discussion

We examined the *what* and *where* of infants' and young children's everyday experiences in the home by characterizing the activities and physical locations in their egocentric view. To do so, we leveraged VideoQA models to annotate a large, naturalistic dataset of egocentric video recordings [23]. While prior work has highlighted the importance of activity contexts in guiding early learning [25], [30], the field has been limited by the sparsity of naturalistic recordings of infant experience and the difficulties in annotating these recordings at scale.

Here, we quantify the consistency and variability in children's everyday activities across physical locations, time, and in their semantic content. To our knowledge this is the first application of VideoQA models to developmental egocentric video data. Our findings join a growing body of work quantifying the regularities in children's everyday experience by leveraging advances in machine learning [46]. With innovations in models, researchers have been able to transcribe the speech that children hear [23] and the locations of faces and hands in view [47], and to create tools to quantify aspects of caregiver-child interactions [48]. An incorporation of the regularities across multiple streams will give a fuller picture of contextual regularities in which children's experiences are situated at both a micro and macro level [49], [50].

Indeed, our findings highlight the need to take a broader perspective on what counts as an activity: singular short, activities (e.g., *eating*) may be embedded in larger activity contexts (e.g., mealtime), and multiple activities may be co-occurring: for example, a young child might walk back and forth between the television and the couch, eating a snack and watching TV simultaneously. While our schema was relatively data driven – starting with unconstrained classifications and prior work – our findings suggest that VideoQA models may still be insufficient to capture some salient activities. Future work that expands this schema in annotation streams (i.e., other VideoQA models and more extensive human annotations), in scope (i.e., other activities, and differentiates between adult-centered and child-centered activities [32]), in time (i.e., to overlapping, longer, and automatically segmented chunks [40]) and in diversity (i.e., outside of the home and Western contexts) will help validate their use for broader contexts and establish generalizability.

The sampling assumptions inherent in the BabyView dataset further limit our inferences. The VideoQA model is currently detecting what a child is *seeing*, and therefore not necessarily what a child is *doing*. Though these are often correlated, they are distinct in the degree to which the child's experience is embedded within an activity. In addition, as parents choose to record on days and at times that are most convenient, there may be an over-representation of some of the most common activities like *playing* and *eating* in our sample (especially relative to activities that are more difficult to capture with a head-mounted camera on, such as *nursing*).

Our results make several contributions to our understanding of why children's everyday experience is challenging to use as training data for current neural network models. First, the present results suggest that naturalistic visual experience contains a small set of activities and locations – again with heavy skewness and thus heavy redundancy in the visual input. To build more efficient and human-like models of visual learning, we may need to modify our algorithms to understand how the visual system learns from such redundant visual scenes and activity contexts [51], [52]. In particular, infants learn to infer visual intermediates such as depth, object identity, and motion from raw sensory input [1], [2], [53], [54] and use them for object reasoning and intuitive physics [2], [3], [55]. We suggest that building models that prioritize learning, extracting, and using intermediate visual representations from redundant visual inputs is a promising avenue for future work.

Second, we observed large variability in the temporal durations for activities, with regularities in how these activities transitioned from one to another. In contrast, much of machine learning trains and evaluates models using carefully curated datasets with neat single-activity clips [56]–[59]; in order to imitate child-like learning, modern models will need to expand to accommodate this temporal variability and redundancy across experienced activities. Overall, our findings suggest (and constrain) testable hypotheses for how current algorithms of human learning must improve to learn from child-like input, charting concrete targets for architectures, objectives, and evaluation protocols.

Children develop sophisticated cognition by learning within structured activity contexts, while these same regularities are at present challenging for most machine learning models. Understanding and exploring this tension will result in more child-like – and likely more efficient – models of human cognition.

Acknowledgments and Disclosure of Funding

This work was supported by an NIH R00HD108386 Pathways to Independence Award to B.L. and a Schmidt Futures gift to M.C.F. We gratefully acknowledge the families who contributed to the dataset, and to the members of the Language and Cognition Lab at Stanford and Visual Learning Lab at UC San Diego for their feedback.

References

[1] P. J. Kellman and E. S. Spelke, "Perception of partly occluded objects in infancy," en, *Cognitive Psychology*, vol. 15, no. 4, pp. 483–524, Oct. 1983, ISSN: 00100285. DOI: 10.1016/0010-0285(83)90017-8. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0010028583900178 (visited on 09/05/2025).

[2] E. S. Spelke and K. D. Kinzler, "Core knowledge," en, *Developmental Science*, vol. 10, no. 1, pp. 89–96, Jan. 2007, ISSN: 1363-755X, 1467-7687. DOI: 10.1111/j.1467-7687.2007.00569.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1467-7687.2007.00569.x (visited on 09/05/2025).

[3] R. Baillargeon, "Infants' Physical World," en, *Current Directions in Psychological Science*, vol. 13, no. 3, pp. 89–94, Jun. 2004, ISSN: 0963-7214, 1467-8721. DOI: 10.1111/j.0963-7214.2004.00281.x. [Online]. Available: https://journals.sagepub.com/doi/10.1111/j.0963-7214.2004.00281.x (visited on 09/05/2025).

[4] P. K. Kuhl, "Early language acquisition: Cracking the speech code," en, *Nature Reviews Neuroscience*, vol. 5, no. 11, pp. 831–843, Nov. 2004, ISSN: 1471-003X, 1471-0048. DOI: 10.1038/nrn1533. [Online]. Available: https://www.nature.com/articles/nrn1533 (visited on 09/05/2025).

[5] K. E. Adolph and J. E. Hoch, "Motor Development: Embodied, Embedded, Enculturated, and Enabling," en, *Annual Review of Psychology*, vol. 70, no. 1, pp. 141–164, Jan. 2019, ISSN: 0066-4308, 1545-2085. DOI: 10.1146/annurev-psych-010418-102836. [Online]. Available: https://www.annualreviews.org/doi/10.1146/annurev-psych-010418-102836 (visited on 09/05/2025).

[6] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," en, *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8619–8624, Jun. 2014, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1403112111. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.1403112111 (visited on 01/29/2024).

[7] S.-M. Khaligh-Razavi and N. Kriegeskorte, "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation," en, *PLoS Computational Biology*, vol. 10, no. 11, J. Diedrichsen, Ed., e1003915, Nov. 2014, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003915. [Online]. Available: https://dx.plos.org/10.1371/journal.pcbi.1003915 (visited on 04/10/2024).

[8] S. Jain and A. G. Huth, *Incorporating Context into Language Encoding Models for fMRI*, en, May 2018. DOI: 10.1101/327601. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/327601 (visited on 06/20/2024).

[9] M. Toneva and L. Wehbe, *Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)*, en, arXiv:1905.11833 [cs, q-bio], Nov. 2019. [Online]. Available: http://arxiv.org/abs/1905.11833 (visited on 03/22/2023).

[10] M. Schrimpf, I. A. Blank, G. Tuckute, *et al.*, "The neural architecture of language: Integrative modeling converges on predictive processing," en, *Proceedings of the National Academy of Sciences*, vol. 118, no. 45, e2105646118, Nov. 2021, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2105646118. [Online]. Available: https://pnas.org/doi/full/10.1073/pnas.2105646118 (visited on 03/22/2023).

[11] C. Caucheteux and J.-R. King, "Brains and algorithms partially converge in natural language processing," en, *Communications Biology*, vol. 5, no. 1, p. 134, Feb. 2022, ISSN: 2399-3642. DOI: 10.1038/s42003-022-03036-1. [Online]. Available: https://www.nature.com/articles/s42003-022-03036-1 (visited on 09/19/2023).

[12] K. L. Aw and M. Toneva, *Training language models to summarize narratives improves brain alignment*, arXiv:2212.10898 [cs, q-bio], Feb. 2023. [Online]. Available: http://arxiv.org/abs/2212.10898 (visited on 05/01/2023).

[13] R. Antonello, A. Vaidya, and A. G. Huth, *Scaling laws for language encoding models in fMRI*, arXiv:2305.11863 [cs], May 2023. [Online]. Available: http://arxiv.org/abs/2305.11863 (visited on 05/22/2023).

[14] K. L. Aw, S. Montariol, B. AlKhamissi, M. Schrimpf, and A. Bosselut, *Instruction-tuning Aligns LLMs to the Human Brain*, arXiv:2312.00575 [cs], Dec. 2023. [Online]. Available: http://arxiv.org/abs/2312.00575 (visited on 12/07/2023).

[15] T. B. Brown, B. Mann, N. Ryder, *et al.*, *Language Models are Few-Shot Learners*, arXiv:2005.14165 [cs], Jul. 2020. DOI: 10.48550/arXiv.2005.14165. [Online]. Available: http://arxiv.org/abs/2005.14165 (visited on 09/05/2025).

[16] A. Radford, J. W. Kim, C. Hallacy, *et al.*, *Learning Transferable Visual Models From Natural Language Supervision*, arXiv:2103.00020 [cs], Feb. 2021. DOI: 10.48550/arXiv.2103.00020. [Online]. Available: http://arxiv.org/abs/2103.00020 (visited on 09/05/2025).

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, arXiv:2010.11929 [cs], Jun. 2021. DOI: 10.48550/arXiv.2010.11929. [Online]. Available: http://arxiv.org/abs/2010.11929 (visited on 09/05/2025).

[18] A. E. Orhan, V. V. Gupta, and B. M. Lake, *Self-supervised learning through the eyes of a child*, arXiv:2007.16189 [cs], Dec. 2020. [Online]. Available: http://arxiv.org/abs/2007.16189 (visited on 08/10/2024).

[19] C. Zhuang, S. Yan, A. Nayebi, *et al.*, "Unsupervised neural network models of the ventral visual stream," en, *Proceedings of the National Academy of Sciences*, vol. 118, no. 3, e2014196118, Jan. 2021, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2014196118. (visited on 08/01/2024).

[20] S. Sheybani, H. Hansaria, J. N. Wood, L. B. Smith, and Z. Tiganj, "Curriculum Learning with Infant Egocentric Videos," en, 2023.

[21] W. K. Vong, W. Wang, A. E. Orhan, and B. M. Lake, "Grounded language acquisition through the eyes and ears of a single child," en, *Science*, vol. 383, no. 6682, pp. 504–511, Feb. 2024, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.adi1374. [Online]. Available: https://www.science.org/doi/10.1126/science.adi1374 (visited on 06/15/2025).

[22] A. E. Orhan, W. Wang, A. N. Wang, M. Ren, and B. M. Lake, *Self-supervised learning of video representations from a child's perspective*, arXiv:2402.00300 [cs, q-bio], Jul. 2024. [Online]. Available: http://arxiv.org/abs/2402.00300 (visited on 08/12/2024).

[23] B. Long, R. Z. Sparks, V. Xiang, *et al.*, "The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences," *Proceedings of the Cognitive Computational Neuroscience Society*, 2025.

[24] M. C. Frank, "Bridging the data gap between children and large language models," en, *Trends in Cognitive Sciences*, vol. 27, no. 11, pp. 990–992, Nov. 2023, ISSN: 13646613. DOI: 10.1016/j.tics.2023.08.007. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1364661323002036 (visited on 09/05/2025).

[25] J. Bruner, "The role of interaction formats in language acquisition," in *Language and social situations*, Springer, 1985, pp. 31–46.

[26] K. Nelson, *Making sense: The acquisition of shared meaning*. New York: Academic Press, 1985.

[27] E. M. Clerkin and L. B. Smith, "Real-world statistics at two timescales and a mechanism for infant learning of object names," *Proceedings of the National Academy of Sciences*, vol. 119, no. 18, e2123239119, 2022.

[28] O. Lavi-Rotbain and I. Arnon, "Visual statistical learning is facilitated in zipfian distributions," *Cognition*, vol. 206, p. 104 492, 2021.

[29] L. Wolters, O. Lavi-Rotbain, and I. Arnon, "Zipfian distributions facilitate children's learning of novel word-referent mappings," *Cognition*, vol. 253, p. 105 932, 2024.

[30] B. C. Roy, M. C. Frank, P. DeCamp, M. Miller, and D. Roy, "Predicting the birth of a spoken word," *Proceedings of the National Academy of Sciences*, vol. 112, no. 41, pp. 12 663–12 668, Oct. 2015. DOI: 10.1073/pnas.1419773112. (visited on 05/20/2022).

[31] T. A. Chang and B. Bergen, "Does Contextual Diversity Hinder Early Word Acquisition?" *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 44, no. 44, 2022. (visited on 09/04/2025).

[32] J. Y. Bang, A. Mora, M. Munévar, A. Fernald, and V. A. Marchman, "Time to talk: Variability in caregiver-child verbal engagement during everyday activities sampled from daylong recordings," *Infancy*, vol. 30, no. 6, e70051, 2025.

[33] M. Soderstrom and K. Wittebolle, "When do caregivers talk? the influences of activity and time of day on caregiver speech and child vocalizations in two childcare environments," *PloS one*, vol. 8, no. 11, e80646, 2013.

[34] E. Hoff-Ginsberg, "Mother-child conversation in different social classes and communicative settings," *Child Development*, vol. 62, no. 4, pp. 782–796, 1991, ISSN: 0009-3920. DOI: 10.2307/1131177. JSTOR: 1131177. (visited on 11/06/2024).

[35] C. S. Tamis-LeMonda, S. Custode, Y. Kuchirko, K. Escobar, and T. Lo, "Routine language: Speech directed to infants during home activities," *Child Development*, vol. 90, no. 6, pp. 2135–2152, 2019, ISSN: 1467-8624. DOI: 10.1111/cdev.13089. (visited on 01/19/2024).

[36] C. R. Rosemberg, F. Alam, M. L. Ramirez, and M. I. Ibañez, "Activity contexts and child-directed speech in socioeconomically diverse argentinian households," *International Journal of Early Childhood*, vol. 55, no. 1, pp. 1–25, 2023.

[37] C. E. Snow and D. E. Beals, "Mealtime talk that supports literacy development," *New directions for child and adolescent development*, vol. 2006, no. 111, pp. 51–66, 2006.

[38] L. Glas, C. Rossi, R. Hamdi-Sultan, C. Batailler, and H. Bellemmouche, "Activity types and child-directed speech: A comparison between french, tunisian arabic and english," *Canadian Journal of Linguistics/Revue canadienne de linguistique*, vol. 63, no. 4, pp. 633–666, 2018.

[39] R. Z. Sparks, B. Long, G. E. Keene, *et al.*, "Characterizing contextual variation in children's preschool language environment using naturalistic egocentric videos," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 46, 2024.

[40] N.-P. Suffo, P.-E. Martin, A. Suffo, D. Haun, and M. Bohn, "Childlens: An egocentric video dataset for activity analysis in children," *PsyArXiv preprint*, 2025. [Online]. Available: https://osf.io/preprints/psyarxiv/evkrf_v1.

[41] B. Zhang, K. Li, Z. Cheng, *et al.*, "VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding," *arXiv preprint arXiv:2501.13106*, 2025. [Online]. Available: https://arxiv.org/abs/2501.13106.

[42] N. Joshi, S. B. Kang, C. L. Zitnick, and R. Szeliski, "Image deblurring using inertial measurement sensors," in *ACM SIGGRAPH 2010 Papers*, Los Angeles, California: Association for Computing Machinery, 2010, ISBN: 9781450302104. DOI: 10.1145/1833349.1778767. [Online]. Available: https://doi.org/10.1145/1833349.1778767.

[43] A. Karpenko, D. Jacobs, J. Baek, and M. Levoy, "Digital video stabilization and rolling shutter correction using gyroscopes," Stanford University, CSTR 2011-03, 2011.

[44] B. Joshi, M. Xanthidis, S. Rahman, and I. Rekleitis, "High definition, inexpensive, underwater mapping," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 1113–1121. DOI: 10.1109/ICRA46639.2022.9811695.

[45] J. Sullivan, M. Mei, A. Perfors, E. Wojcik, and M. C. Frank, "SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective," *Open mind*, vol. 5, pp. 20–29, 2021.

[46] M. C. Frank and N. D. Goodman, "Cognitive modeling using artificial intelligence," *PsyArXiv. Retrieved from osf. io/preprints/psyarxiv/wv7mg v1 doi*, vol. 10, 2025.

[47] B. L. Long, G. Kachergis, K. Agrawal, and M. C. Frank, "A longitudinal analysis of the social information in infants' naturalistic visual experience using automated detections.," *Developmental Psychology*, vol. 58, no. 12, p. 2211, 2022.

[48] Z. Weng, L. Bravo-Sánchez, Z. Wang, *et al.*, "Artificial intelligence–powered 3d analysis of video-based caregiver-child interactions," *Science Advances*, vol. 11, no. 8, eadp4422, 2025.

[49] M. L. Rowe and A. Weisleder, "Language development in context," *Annual Review of Developmental Psychology*, vol. 2, no. 1, pp. 201–223, 2020.

[50] M. Casillas, "Learning language in vivo," *Child Development Perspectives*, vol. 17, no. 1, pp. 10–17, 2023.

[51] D. M. Bear, K. Feigelis, H. Chen, *et al.*, *Unifying (Machine) Vision via Counterfactual World Modeling*, arXiv:2306.01828 [cs], Jun. 2023. [Online]. Available: http://arxiv.org/abs/2306.01828 (visited on 04/04/2024).

[52] A. Bardes, Q. Garrido, J. Ponce, *et al.*, *Revisiting Feature Prediction for Learning Visual Representations from Video*, arXiv:2404.08471 [cs], Feb. 2024. [Online]. Available: http://arxiv.org/abs/2404.08471 (visited on 08/14/2024).

[53] E. S. Spelke, "Principles of Object Perception," en, *Cognitive Science*, vol. 14, no. 1, pp. 29–56, Jan. 1990, ISSN: 0364-0213, 1551-6709. DOI: 10.1207/s15516709cog1401_3. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1207/s15516709cog1401_3 (visited on 09/05/2025).

[54] M. Kavšek, A. Yonas, and C. E. Granrud, "Infants' sensitivity to pictorial depth cues: A review and meta-analysis of looking studies," en, *Infant Behavior and Development*, vol. 35, no. 1, pp. 109–128, Feb. 2012, ISSN: 01636383. DOI: 10.1016/j.infbeh.2011.08.003. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0163638311000877 (visited on 09/05/2025).

[55] E. Téglás, E. Vul, V. Girotto, M. Gonzalez, J. B. Tenenbaum, and L. L. Bonatti, "Pure Reasoning in 12-Month-Old Infants as Probabilistic Inference," en, *Science*, vol. 332, no. 6033, pp. 1054–1059, May 2011, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1196404. [Online]. Available: https://www.science.org/doi/10.1126/science.1196404 (visited on 09/05/2025).

[56] K. Soomro, A. R. Zamir, and M. Shah, *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild*, arXiv:1212.0402 [cs], Dec. 2012. DOI: 10.48550/arXiv.1212.0402. [Online]. Available: http://arxiv.org/abs/1212.0402 (visited on 09/05/2025).

[57] W. Kay, J. Carreira, K. Simonyan, *et al.*, *The Kinetics Human Action Video Dataset*, arXiv:1705.06950 [cs], May 2017. DOI: 10.48550/arXiv.1705.06950. [Online]. Available: http://arxiv.org/abs/1705.06950 (visited on 09/05/2025).

[58] R. Goyal, S. E. Kahou, V. Michalski, *et al.*, *The "something something" video database for learning and evaluating visual common sense*, arXiv:1706.04261 [cs], Jun. 2017. DOI: 10.48550/arXiv.1706.04261. [Online]. Available: http://arxiv.org/abs/1706.04261 (visited on 09/05/2025).

[59] M. Monfort, A. Andonian, B. Zhou, *et al.*, *Moments in Time Dataset: One million videos for event understanding*, arXiv:1801.03150 [cs], Feb. 2019. DOI: 10.48550/arXiv.1801.03150. [Online]. Available: http://arxiv.org/abs/1801.03150 (visited on 09/05/2025).

A Appendices

A.1 Additional annotation procedure details

The exact prompt we used to generate activities, locations, and video descriptions from VideoLLaMA3 was "This video is recorded from the point-of-view of a child, with a camera mounted on the child's head. Respond strictly only in this format with both keys and values: Location: <balcony/bathroom/bedroom/car/closet/deck/porch/dining room/garage/hallway/kitchen/laundry room/living room/office/outside/playroom/ stairway/storage room/other> || Activity: <being held/cleaning/cooking/conversing/crawling/crying/drawing/drinking/eating/exploring/gardening/getting dressed/looking at device/lying down/music/nursing/overhearing speech/playing/reading/sitting/standing/walking/other> || Video description: ... ||"

A.2 Additional annotation validation details

Each coder was provided with the same 100 sampled clips to annotate. 50 of these clips were randomly sampled from the dataset. The other 50 were selected from a subset of the dataset with uniform coverage of each existing activity and location pair, to ensure that we were validating our model across frequent and infrequent activity and location detections.

A manual inspection of a subset of the clips by authors suggested that another key reason why the activity detection task is non-trivial for both human raters and the VideoQA model, other than multiple activities potentially co-occurring within a single clip, is that activities are semantically similar to each other (e.g., *watching TV* and *looking at device*), and some clips could very reasonably have

two similar annotations. Location annotations, while more reliable, also benefited from including additional locations, accounting for videos where the child moved between locations within a 10-second duration (location count M=1.19, SD=0.41). Upon including these locations as alternative ground truths, location precision increased to 0.69 and increased further to 0.82 when collapsing across annotators.

A.3 Additional inertial measurement unit details

Accelerometer (translation). The accelerometer responds to *translation*—how the device's position in the world changes over time. In practice, it measures a combination of true linear acceleration (e.g., speeding up, slowing down, or changing direction) and the ever-present pull of Earth's gravity. If you hold the device still, the accelerometer still reports gravity; if you move it straight up or forward, it records the extra push from that translational motion. We first remove the effect of gravity by subtracting the device's gravity estimate from the raw accelerometer signal, leaving *linear acceleration* due to actual movement. We then express this movement relative to gravity so the results make sense regardless of head (and camera) orientation.

Gyroscope (rotation). The gyroscope, in contrast, measures *rotation*: how fast and about which axis the device is turning (angular velocity). Yaw, pitch, and roll are different mixtures of the three measured axes; nodding, shaking, or tilting the head produces distinctive gyroscope signals even if the device is not translating. Together, these sensors let us disentangle whether a change in the camera's viewpoint came from the head *moving through space* (translation, accelerometer) or *turning in place* (rotation, gyroscope), which is crucial for interpreting head motion during everyday activities. In practice, because the sensor is head-mounted, head rotations can also induce small linear accelerations (e.g., due to the sensor's offset from the neck pivot). we compute overall rotational activity by taking the Euclidean (ℓ_2) norm of angular velocity across the three axes – a single total rotation speed. This gives a clean, direction-agnostic measure of how much the head is turning, nodding, or tilting at each moment.

A.4 Activity and location prevalence and filtering

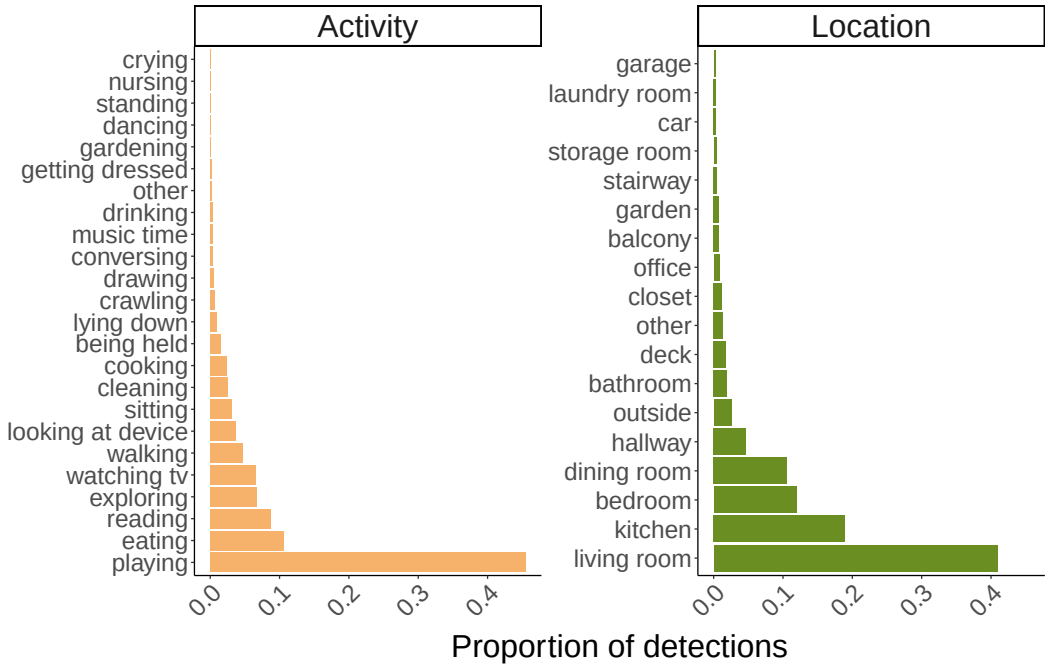

Figure 8: **Prevalence of activities and locations in the dataset.** Overall proportions of activities and locations in the dataset. We found a skewed distribution of both activities and locations in view.

Figure 8 shows the prevalence of activities and locations in the dataset. We found a skewed distribution of both the activities and locations in the infant view. Some activities and locations were much more prevalent than others. The most common activity, *playing*, occurred in 144,329 of the 312,485 clips (400.91 of 868 hours) while the least common activity, *crying*, occurred in only 21 clips (3.5 minutes). However, the low prevalence of *crying* in particular may be caused by the difficulty involved in detecting *crying* in egocentric videos without including audio, and because parents choose to record at times when it is most convenient. Most recordings took place in the living room, in 129,944 clips (360.95 hours). On the other hand, recordings only took place in the garage, the least common location, in 860 clips (2.4 hours)

For some further analyses, activities were filtered to the eight most frequent non-postural activities: *playing, eating, reading, exploring, watching TV, looking at device, cleaning,* and *cooking*. We defined postural activities as activities that primarily involved an aspect of the child's posture or locomotive ability. We excluded postural activities like *walking* and *sitting* from these analyses since they are more passive and inherently always co-occurring with any present non-postural activities.

A.5 Reading clustering results

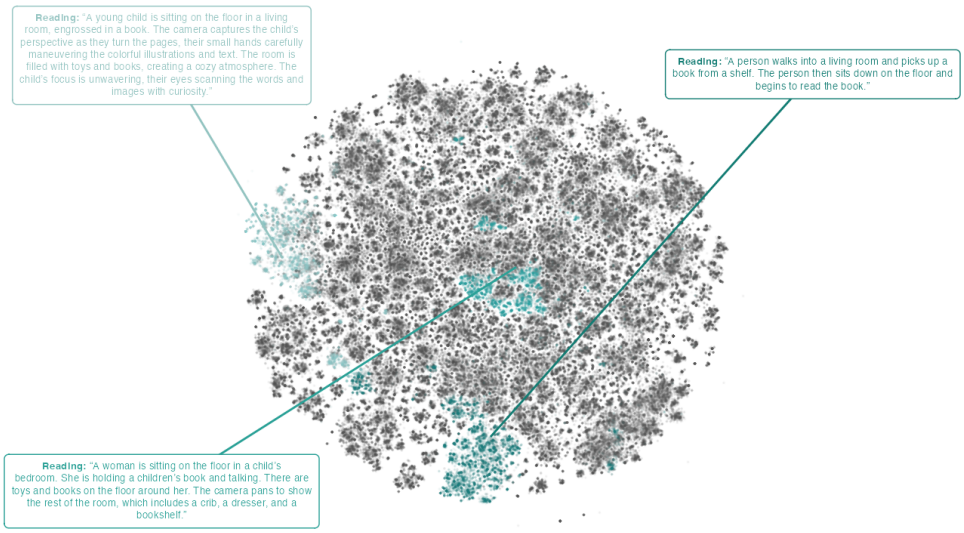

Figure 9: **t-SNE visualization of *reading* description clusters.** 3-means clustering of video description embeddings revealed distinct subtypes of reading, colored by cluster. Grey dots indicate embeddings for all activities other than reading. Text labels provide example descriptions for different activities.

Figure 9 shows the t-SNE plot for the three different clusters, clustered using 3-means clustering. The text descriptions (as well as visual inspection) revealed different subtypes of reading, namely reading with the child close to the book (light teal), reading with the child far from the book (dark teal), and more ambiguous cases (medium teal).